RETHINKING MULTI-SOURCE PROMPT TRANSFER: A COMMON-KNOWLEDGE-GUIDED ORTHOGONAL ADAPTATION STRATEGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-source soft prompt transfer reuses prompts learned on existing tasks to provide shared knowledge and complementary information for a target task. However, the optimality of common-knowledge extraction and the relationship between common and task-specific knowledge remain insufficiently understood. We revisit this process from the perspectives of parameter reconstruction risk and optimization geometry and propose COPT, a common-knowledge-guided orthogonal prompt transfer method. Under assumptions of knowledge representability, weighted centering of task-specific offsets, and squared Frobenius reconstruction risk, we show that a truncated singular value decomposition of the mean source prompt yields an optimal common prompt subject to a rank constraint. We then construct a shared subspace from this common prompt and project gradients onto its orthogonal complement during source-specific knowledge learning, suppressing gradient leakage into shared directions. Finally, we utilize attention mechanisms to integrate general and domain-specific knowledge, enabling adaptation to target tasks. With 77K trainable parameters per target task on T5-Base, COPT achieves average scores of 87.00 on GLUE and 78.14 on SuperGLUE, outperforms the SOTA baseline. Controlled preliminary experiments, ablations, and analyses of hyperparameters and model scale further support the effectiveness of common-knowledge extraction and source-side orthogonal constraints.

1 INTRODUCTION

Background. Compared with full fine-tuning, prompt tuning (PT) prepends a set of trainable continuous vectors to the input and has been shown to achieve comparable performance on various downstream NLP tasks at relatively low cost (Lester et al., 2021), making it a promising approach to model adaptation. Conventional soft prompt tuning, however, typically relies on random initialization, which can lead to unstable training and underuse of knowledge learned on existing tasks. To address this issue, recent research has focused on soft prompt transfer: reusing prompt parameters learned on source tasks to provide more informative initialization or prior knowledge for a target task, which has been shown to significantly improve prompt-tuning performance (Vu et al., 2022). As reusable prompt libraries grow, jointly exploiting prompts from multiple sources has become a key focus in the field of soft prompt transfer.

Shortcomings. The aforementioned concepts have spurred research into multi-source prompt transfer. As illustrated in Figure 1, the mainstream approaches rely on the extraction of shared and specific knowledge followed by adaptation to the target task. For instance, MPT extracts transferable shared knowledge via knowledge distillation and achieves adaptation to target tasks by incorporating task-specific updates (Wang et al., 2023). PCT further extracts task-agnostic representations from source task prompts and enabling instance-level knowledge selection and composition for the target task through contrastive learning and attention mechanisms (Zhao et al., 2025). A detailed review of soft prompt tuning and prompt-transfer methods is provided in Appendix C. Although effective, there are still two aspects that haven't been fully explored: on one hand, common knowledge is usually extracted based on empirical constructs and lacks a unified theoretical framework; on the other hand, the training process for specialized knowledge typically lacks consideration of intru-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

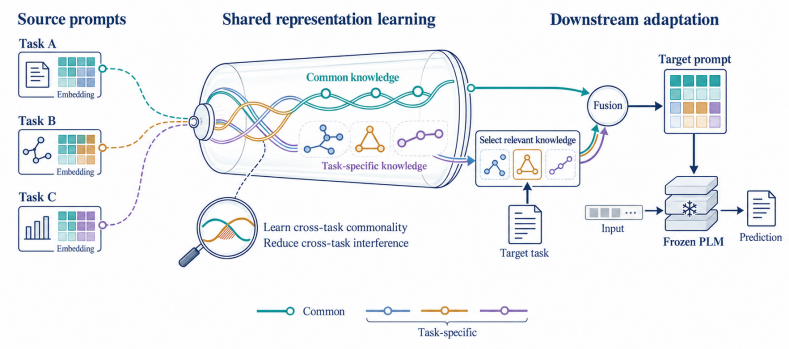


Figure 1: Multi-source prompt transfer through common-knowledge extraction and task-specific knowledge selection. COPT uses common knowledge to guide source-specific learning and combines relevant source information for downstream adaptation with a frozen language model.

sions into the common knowledge space, which may lead to knowledge overlap and affect transfer performance.

Motivation. Therefore, we revisit multi-source soft prompt transfer and make the following observations: (i) under the squared Frobenius risk and a centering assumption, common-knowledge extraction admits a theoretically optimal solution; and (ii) constraining task-specific updates to the orthogonal complement of the common-knowledge subspace reduces their overlap in parameter space and improves the resulting transfer performance.

Proposed Solution. Based on these observations, we propose COPT, a common-knowledge-guided orthogonal prompt transfer method. Specifically, based on the established theory, we first average source soft prompts and apply SVD to construct a low-rank common prompt that captures shared knowledge. We then use this prompt as a reference to construct a shared subspace and introduce trainable matrices to represent task-specific knowledge. During training, we project gradients onto the orthogonal complement of the shared subspace to suppress leakage of task-specific information into common directions. Finally, COPT uses attention mechanism to select task-specific representations relevant to the target task and assign them weights. The weighted task-specific embeddings are combined with the task-independent embedding to form a target prompt, enabling input-dependent knowledge transfer.

Contributions. Our main contributions are:

- **Re-examining multi-source prompt transfer**. Our work establishes a feasible theoretical framework for common knowledge, proposing a theoretically optimal scheme for acquiring common knowledge under weighted centering and squared Frobenius risk; simultaneously, it identifies how existing methods encroach upon the common knowledge space during task-specific knowledge training.
- **A more efficient prompt transfer method**. Based on these findings, we propose COPT, a common-knowledge-guided orthogonal prompt transfer method. By employing a novel common knowledge extraction approach alongside orthogonality constraints relative to task-specific knowledge, we achieve efficient multi-source prompt transfer. Experimental results demonstrate that our method outperforms the baselines, achieving improvements of 0.80 and 0.64 percentage points over PCT on GLUE and SuperGLUE, respectively; ablation and preliminary experiments further validate the feasibility of the underlying theory and component designs.

2 Rethinking Multi-Source Soft Prompt Transfer

2.1 Rethinking Common-Knowledge Extraction

We first focus on common knowledge extraction. Existing approaches extract common knowledge in different ways, including matrix concatenation and low-rank reconstruction (Wang et al., 2023; Zhao et al., 2025). A shared limitation is the lack of a unified theoretical framework for comparing

these choices. We first state the underlying assumptions, then derive an optimality result for common knowledge in terms of parameter risk and examine it empirically.

Assumption 3.1 Knowledge Representability and Unbiasedness

Assume that each source prompt decomposes into a common-knowledge matrix C_0 and a task-specific offset T_i, such that $P_i = C_0 + T_i$. For fixed positive weights $w_i > 0$. Further assume that the task-specific offsets in the current finite set of sources satisfy weighted centering: $\sum_{i=1}^{n} w_i T_i = 0$

Assumption 3.2 Knowledge Comparability through Frobenius Distance

Under Assumption 3.1, let $\mathcal{C} \subseteq \mathbb{R}^{m \times d}$ be a nonempty closed candidate set. Our goal is to find the candidate that retains the most common knowledge. To compare knowledge content across matrices, we introduce the squared Frobenius distance. This norm is commonly used in matrix compression, where the norm of the difference between two matrices measures reconstruction loss. We assume that a candidate common prompt retains more common knowledge at the parameter-representation level when it reconstructs C_0 more accurately. We therefore measure common-knowledge reconstruction loss by $\|C - C_0\|_F^2$, providing a unified criterion for comparing candidate prompts.

Definition 1 Optimal Common Prompt

Under Assumption 3.2, extracting common knowledge amounts to minimizing the reconstruction loss between a candidate matrix and C_0. We accordingly define an optimal common prompt as

$$C_{\mathcal{C}}^{*} \in \arg\min_{C \in \mathcal{C}} \|C - C_0\|_F^2. \tag{1}$$

Theorem 1 Equivalent Characterization of the Optimal Common Prompt

Combining the assumptions, we can derive an equivalent expression for the optimal public prompt. The proof is provided in Appendix A.1.

$$\arg\min_{C \in \mathcal{C}} \|C - C_0\|_F^2 = \arg\min_{C \in \mathcal{C}} \sum_{i=1}^{n} w_i \|P_i - C\|_F^2. \tag{2}$$

Theorem 2 Optimality of Truncated SVD for a Low-Rank Common Prompt

Further impose rank constraints, let $\mathcal{C}_r = \{C \in \mathbb{R}^{m \times d} : \operatorname{rank}(C) \leq r\}$ for $0 \leq r \leq q = \min(m, d)$. By the EckartYoungMirsky theoremEckart & Young (1936), truncated SVD gives a global minimizer of Equation 2. Let $\overline{P}_w = U \Sigma V^{\top}$ be the SVD of the weighted average of source prompts. Then the optimal low-rank common prompt is

$$C_r^{*} = [\overline{P}_w]_r = \sum_{j=1}^{r} \sigma_j u_j v_j^{\top}. \tag{3}$$

This solution minimizes common-knowledge reconstruction loss subject to the rank constraint, providing a theoretical basis for constructing a low-rank common prompt. The proof is given in Appendix A.3.

2.2 Rethinking Task-Specific Knowledge Learning

The acquisition of task-specific knowledge occurs after the extraction of common knowledge, by further training to isolate information relevant to specific tasks from various source prompts, which is then used for the selective integration in subsequent target tasks. Meanwhile, we note that in multi-task learning(Wei et al., 2025; Yu et al., 2020), there are also methods that utilize knowledge decomposition and routing to achieve transfer learning on target tasks. Related studies indicate that the update direction of knowledge training gradients is an important factor affecting performance. However, the aforementioned multi-source soft prompt methods, after fixing the public prompt, have not yet imposed explicit constraints on the update direction of task-specific knowledge, which may lead to an overlap between the two in the parameter space and further affect performance. **To test**

this hypothesis, this paper defines a public subspace leakage rate to quantify the encroachment of task-specific knowledge on the public space, and introduces an orthogonal complement constraint as a solution, finally validated through experiments.

Public Subspace Leakage Rate. Let a trainable residual $D_i \in \mathbb{R}^{m\times d}$ represent source-specific knowledge. We construct an orthonormal basis $B \in \mathbb{R}^{m\times r_b}$ along the token dimension from the common prompt, satisfying $B^\top B = I_{r_b}$ and $0 < r_b < m$, and keep it fixed during training. The common subspace is $span(B)$. For a training example z, let the loss be $\updownarrow_i(D_i; z)$ and its gradient be $g = \nabla_{D_i}\updownarrow_i(D_i; z)$. When $B^\top g \neq 0$, the gradient for task-specific learning contains a component in the common directions. To measure the overlap between task-specific residuals and common directions, we define the common-subspace leakage ratio as

$$\rho_{param}(D_i) = \frac{\|BB^\top D_i\|_F^2}{\|D_i\|_F^2 + \varepsilon}, \varepsilon > 0. \tag{4}$$

Here, $BB^\top$ is the projector onto the common subspace, and ε is a numerical stability constant. A larger ρ_{param} indicates that more of the residual's relative energy lies in common directions, corresponding to greater overlap between common and task-specific knowledge in parameter space.

orthogonal complement constraint. To improve the distinguishability of common and task-specific knowledge, we introduce an orthogonal-complement constraint and decompose the training gradient into a common-subspace component and an orthogonal component:

$$g = g_\| + g_\perp, g_\| = BB^\top g, g_\perp = (I_m - BB^\top)g. \tag{5}$$

Here, $g_\|$ corresponds to common directions and $g_\perp$ to their orthogonal complement. Only $g_\perp$ is passed to the optimizer. Since $B^\top B = I_{r_b}$, we have $B^\top g_\perp = 0$, eliminating the common-direction component of the gradient. The loss and gradient-update procedure are developed further in the method section.

2.3 Empirical Verification

The preceding analysis yields two testable implications: mean-prompt SVD should improve the extraction of common knowledge, and restricting source-specific updates to the orthogonal complement of the common subspace should reduce redundancy. To examine these implications under practical transfer conditions, we use PCT (Zhao et al., 2025) as the common baseline and conduct two controlled comparisons. In the first comparison, we replace PCT's common-knowledge extraction procedure with the proposed mean-prompt procedure while leaving the subsequent transfer process unchanged (Theory-PCT). In the second, we retain PCT's extraction procedure and apply the orthogonal-complement projection only to the gradients used for source-specific knowledge learning (Orthogonal-PCT). Table 1 summarizes the average GLUE performance and the common-subspace leakage ratio; the corresponding task-level results and experimental settings are given in Appendices A.4 and A.5.

Table 1: Controlled preliminary experiments on common-knowledge extraction and orthogonal-complement constraints.

Metric	PCT	Theory-PCT	Orthogonal-PCT
GLUE Avg.	80.49	81.63 (+1.14)	83.90 (+3.41)
ρ_{param}	0.1900	-	0.1519 (-0.0381)

The two comparisons support the proposed design. Replacing PCT's extraction procedure with mean-prompt SVD raises the average GLUE score from 80.49 to 81.63, an improvement of 1.14 percentage points, providing empirical support for the reconstruction-based common-knowledge criterion. Applying the orthogonal-complement update increases the average score further to 83.90 and reduces the leakage ratio from 0.1900 to 0.1519, corresponding to an approximately 20.1% relative reduction. The result links the reduction in common-space overlap to the observed transfer

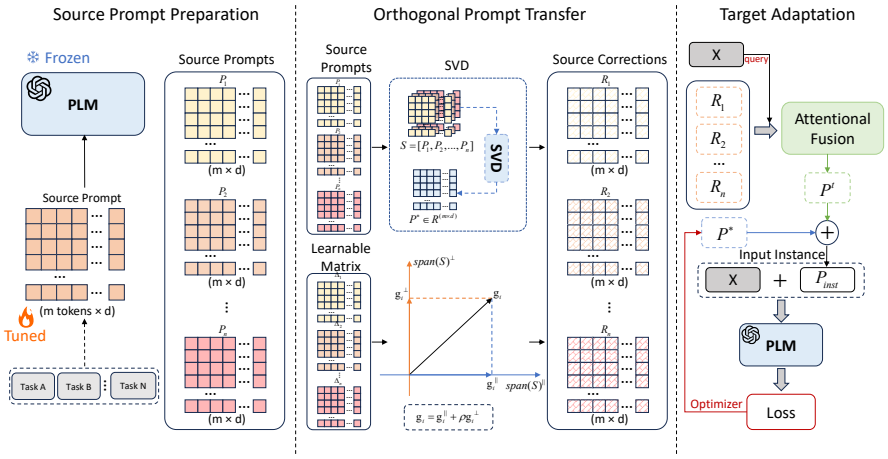


Figure 2: COPT extracts common knowledge based on optimality theory and orthogonally separates task-specific knowledge from it, thereby reducing knowledge overlap and task conflict while enhancing the efficiency and performance of multi-source prompt transfer.

improvement and motivate combining theoretically guided extraction with directional control in COPT.

3 METHOD

3.1 OVERVIEW

Section 2 established how to estimate common knowledge and why task-specific updates should be separated from it. We now turn these findings into COPT. As illustrated in Figure 2, COPT implements a three-stage transfer procedure: Extract common knowledge by SVD; Learn source-specific transformations with projected gradients; Adapt a target transformation using input-dependent fusion.

3.2 COMMON-KNOWLEDGE EXTRACTION

Common-knowledge extraction seeks task-independent embeddings from the collection of source prompts. > Consider n high-resource source tasks, each associated with a soft prompt $P_i \in \mathbb{R}^{m\times d}$, where m is the prompt length and d is the pretrained model's hidden dimension. As established in Section 2.1, truncated SVD of the average of the source prompts yields an optimal low-rank estimate of common knowledge. We extract common knowledge based on this theory.

$$\begin{aligned} \overline{P} &= \frac{1}{n}\sum_{i=1}^{n} P_i, \\ SVD(\overline{P}) &= \overline{U}\overline{\Sigma}\overline{V}^{\top} \in \mathbb{R}^{m\times d}. \end{aligned} \tag{6}$$

Retaining the first r_p components gives the common prompt C^*:

$$C^* = \overline{U}_{[:,1:r_p]}\overline{\Sigma}_{1:r_p,1:r_p}\overline{V}^{\top}_{[:,1:r_p]} \in \mathbb{R}^{m\times d}. \tag{7}$$

3.3 TASK-SPECIFIC KNOWLEDGE LEARNING

After obtaining the common prompt, we separate source-specific embedding information from each source prompt for subsequent target-task adaptation. Following the analysis in Section 3, we firstly

construct a shared subspace from the common knowledge and project training gradients onto its orthogonal complement to better distinguish task-specific knowledge from common knowledge.

3.3.1 SHARED SUBSPACE CONSTRUCTION

The shared subspace is spanned by the leading left singular vectors of C^*.

$$C^* = U^*\Sigma^*(V^*)^\top, B = U^*_{[:,1:r_b]} \in \mathbb{R}^{m\times r_b}. \tag{8}$$

Based on this, the projection operator on the shared subspace and its orthogonal complement can be expressed as:

$$\Pi_{\|}(G) = BB^\top G, \Pi_{\perp}(G) = (I_m - BB^\top)G, G \in \mathbb{R}^{m\times d}. \tag{9}$$

$\Pi_{\|}(G)$ retains components aligned with common token directions, whereas $\Pi_{\perp}(G)$ removes them and supplies the update used for source-specific learning. This gives orthogonal adaptation a direct operational meaning.

3.3.2 GRADIENT PROJECTION AND TASK-SPECIFIC OPTIMIZATION

For each source task, we introduce a trainable parameter matrix $M_i \in \mathbb{R}^{m\times d}$ of the same shape as the source prompt to store task-specific embeddings. Each M_i combines with common knowledge through a Hadamard product to reconstruct the source prompt as $Q_i = M_i \odot C^*$. Specialized knowledge training aims at reconstruction loss to ensure the maximal preservation of source prompt knowledge:

$$\mathcal{L}_{rec} = \frac{1}{n}\sum_{i=1}^{n} \|Q_i - P_i\|_F^2. \tag{10}$$

Let $G_i = \nabla_{M_i}\mathcal{L}_{rec}$ be the original gradient for source task i. Before each parameter update, COPT applies the orthogonal-complement projection using the shared basis:

$$\widetilde{G}_i = \Pi_{\perp}(G_i) = (I_m - BB^\top)G_i, B^\top\widetilde{G}_i = 0. \tag{11}$$

The optimizer receives only $\widetilde{G}_i$ and updates M_i based on $M_i \leftarrow M_i - \eta\widetilde{G}_i$.

3.4 TARGET-TASK ADAPTATION

In the process of adapting to the target task, we integrate general knowledge with task-specific knowledge to generate concrete prompts tailored to the target task. First, the attention mechanism is employed to determine the weight of each type of knowledge in the final prompt. Then, a learnable embedding matrix is used to adapt the target prompt content.

Attention-Based Weighting of Task-Specific Knowledge. Following Zhao et al. (Zhao et al., 2025), we determine the contribution of each task-specific embedding from the target input. For a target input x, let its token embeddings be $E(x)$. We compute the inner product between the target-input representation and each task-specific embedding representation to obtain attention weights, applying a softmax with temperature T to avoid overly confident assignments:

$$\alpha_j(x) = \frac{\exp(E(x)^\top M_j/T)}{\sum_{h=1}^{n}\exp(E(x)^\top M_h/T)}, j = 1,\ldots,n. \tag{12}$$

Target Adaptation Training. Then, COPT introduces a learnable target-side embedding $M_t \in \mathbb{R}^{m\times d}$ and combines the common knowledge, task-specific embeddings, and attention weights to obtain the target prompt:

$$M_{inst}(x) = M_t + \sum_{j=1}^{n} \alpha_j(x) M_j, Q_t(x) = M_{inst}(x) \odot C^* \in \mathbb{R}^{m \times d}. \tag{13}$$

Finally, update M_t by supervised training to adapt to the target task.

$$D^{(t)*} \in \underset{D^{(t)} \in \mathbb{R}^{m \times d}}{\arg\min} - \frac{1}{N_t} \sum_{j=1}^{N_t} \log p_\theta \left(y_j \mid M_{inst}(x); E(x_j) \right). \tag{14}$$

Common and source-specific knowledge are provided by the frozen C^* and $\{M_i\}$, so only the target-side embedding M_t is trained. No source-side common-subspace gradient projection is applied to M_t during target adaptation, allowing optimization to adjust to the target-task distribution.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate COPT on the GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a) natural language understanding benchmarks. For GLUE, we use MNLI, QNLI, QQP, and SST-2 as source tasks. For SuperGLUE, we additionally include ReCoRD and SQuAD to incorporate reading-comprehension knowledge and broaden source-task coverage.

Baselines. We compare against full fine-tuning (Raffel et al., 2020) and several parameter-efficient methods, including Adapter (Houlsby et al., 2019), BitFit (Zaken et al., 2022), LoRA (Hu et al., 2022), and QLoRA (Dettmers et al., 2023). The prompt-tuning baseline is Prompt Tuning (PT) (Lester et al., 2021). Prompt-transfer baselines include similarity-based SPoT (Vu et al., 2022), attention-based ATTEMPT (Asai et al., 2022), multi-task distillation-based MPT (Wang et al., 2023), PANDA (Zhong et al., 2024), which combines prompt transfer with knowledge distillation, DTVG, and PCT (Zhao et al., 2025), which separates common and task-specific knowledge through contrastive transformation.

Evaluation metrics. We follow the standard metric settings for the datasets: Pearson correlation for STS-B, Matthews correlation for CoLA, F1 for MultiRC, and accuracy for the remaining tasks. All results are reported as percentages. GLUE Avg. and SuperGLUE Avg. denote the macro-average across target tasks within each benchmark. Detailed implementation details are provided in Appendix D.

4.2 Main Results

Tables 1 and 2 present the experimental results of various baseline methods across the four benchmarks. On GLUE and SuperGLUE, COPT achieves new state-of-the-art performance. Compared to the basic Prompt Tuning, COPT improves the average GLUE score by 14.80 points and the average SuperGLUE score by 20.34 points. With only 77K trainable parameters per task, COPT outperforms full fine-tuning by 2.10 points on GLUE and 4.24 points on SuperGLUE, demonstrating strong parameter efficiency.

4.3 Ablation Study

To verify the effectiveness of components, we examine three variants on the eight GLUE target tasks: replacing common-prompt construction with PCT's procedure, removing the source-side orthogonal constraint, and applying both changes. We report the average test scores on GLUE.

As shown in Figure 3(a), COPT achieves the highest average score of 86.44. Removing common-knowledge extraction or the orthogonal constraint lowers performance to 85.30 and 83.60, respectively, while removing both lowers it further to 83.20. This indicates that both common-prompt construction and directional constraints on task-specific learning contribute to transfer performance, with their combination achieving the best result.

Table 2: Results on GLUE and SuperGLUE (%). Most of the experimental results come from the studies of ATTEMPT(Asai et al., 2022) and PCT(Zhao et al., 2025), and the experimental setup is consistent with their research. Bold indicates the best result in each task column; underlining indicates the second-best result.

		GLUE									SuperGLUE					
Method	Params/task	MNLI	QQP	QNLI	SST-2	RTE	CoLA	STS-B	MRPC	Avg.	MultiRC	BoolQ	WiC	WSC	CB	Avg.
Fine-Tuning Methods																
Fine Tuning	220M	86.8	**91.6**	93.0	94.6	71.9	61.8	89.7	**90.2**	84.9	72.8	81.1	70.2	59.6	85.7	73.9
Adapter	1.9M	86.5	90.2	93.2	93.8	71.9	**64.0**	90.7	85.3	84.5	75.9	**82.5**	67.1	**67.3**	85.7	75.7
BitFit	280K	85.3	90.1	93.0	94.2	67.6	58.2	90.9	86.8	83.3	74.5	79.6	70.0	59.6	78.6	72.5
LoRA	880K	83.7	86.9	92.0	94.0	79.1	61.3	90.5	87.3	84.4	75.7	80.0	68.7	55.8	**94.6**	75.0
QLoRA	880K	85.8	89.7	93.5	94.3	81.6	61.3	90.6	87.2	85.5	75.6	81.8	67.1	63.5	**94.6**	76.5
PT	77K	81.3	89.7	92.8	90.9	54.7	10.6	89.5	68.1	72.2	58.7	61.7	48.9	51.9	67.9	57.8
Prompt Transfer Methods																
SPoT	77K	85.4	90.1	93.0	93.4	69.8	57.1	90.0	79.7	82.3	74.0	77.2	67.0	50.0	46.4	62.9
ATTEMPT	232K	84.3	90.3	93.0	93.2	73.4	57.4	89.7	85.7	83.4	74.4	78.8	66.8	53.8	78.6	70.5
MPT	78K	85.9	90.3	93.1	93.8	79.4	62.4	90.4	89.1	85.6	74.8	79.6	69.0	**67.3**	79.8	74.1
PANDA	77K	82.2	90.1	93.3	94.2	79.8	45.8	89.8	87.9	82.9	73.4	77.8	64.6	63.5	82.4	72.3
DTVG	77K	85.9	90.3	93.0	93.2	83.2	63.1	91.1	88.3	86.0	74.6	81.5	71.0	66.9	94.5	77.7
PCT	77K	86.1	90.6	**93.8**	94.6	82.3	63.4	90.5	88.3	86.2	**76.0**	82.1	70.2	64.5	**94.6**	77.5
COPT	77K	**87.8**	90.2	93.3	**95.6**	**83.9**	63.5	**91.3**	**90.2**	**87.0**	75.2	81.5	**72.0**	**67.3**	**94.6**	**78.1**

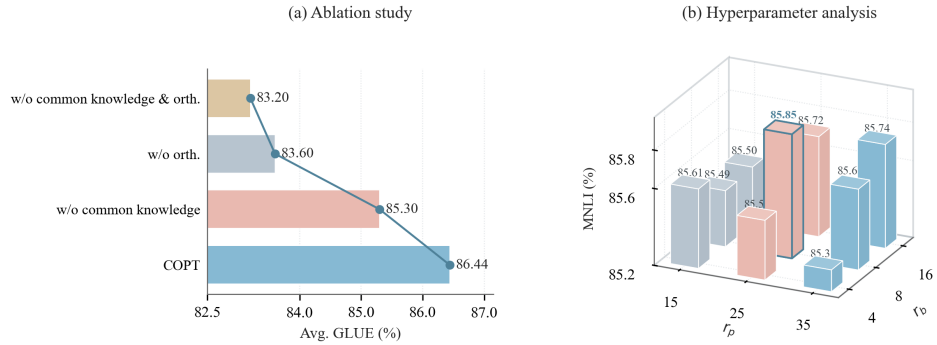


Figure 3: Ablation and hyperparameter analyses. (a) Average GLUE performance for COPT and its variants with common-prompt construction and orthogonal constraints removed. (b) MNLI accuracy for the nine rank combinations.

4.4 Lightweight and Modular Adaptation

Figure 4.4(a) shows that increasing the number of source prompts from three to six raises COPT's RTE score from 78.42 to 82.42 with fluctuation. This suggests that additional sources can provide complementary information, but their benefit depends on how well the transferred knowledge matches the target task.Figure 4.4(b) compares the estimated source-knowledge preparation complexity of three representative methods. Like PCT and other weighted prompt-transformation methods, COPT remains lightweight and modular. In contrust, MPT executes multiple teacher prompts, so its source preparation involves backbone computation and grows with the number of sources. Appendix B provides the detailed estimates and measured running times for COPT.

4.5 Further Analysis

4.5.1 Hyperparameter Analysis

We evaluate the joint effect of the common-prompt rank r_p and shared-basis dimension r_b on MNLI. As shown in Figure 3(b), the nine configurations range from 85.31% to 85.85%, with the best result obtained at $(r_p, r_b) = (25, 8)$, the default setting. The variation across rows indicates that the two ranks should be selected jointly.

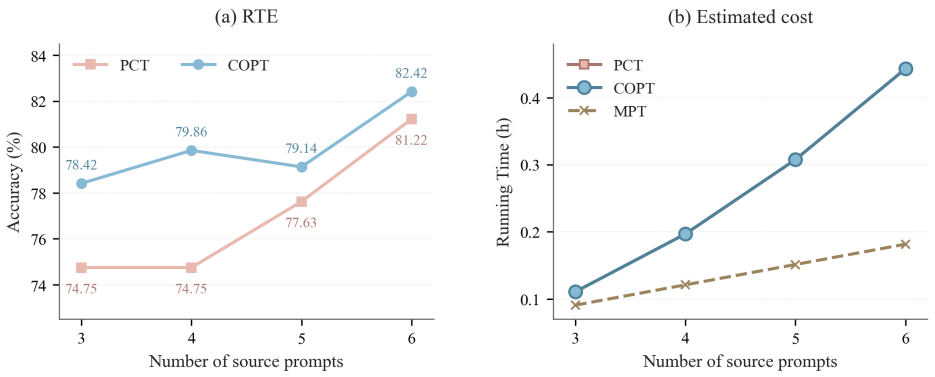


Figure 4: Performance and estimated cost as the source-prompt collection grows in the order MNLI, QNLI, QQP, SST-2, SQuAD, and ReCoRD.

4.5.2 Performance across Model Scales

To examine scale sensitivity, we compare PT, PCT, and COPT on WSC, MultiRC, and WiC with T5-Small, T5-Base, and T5-Large.

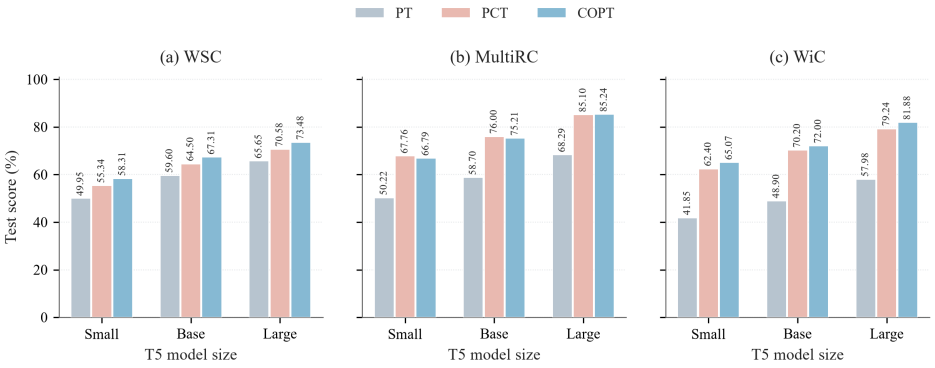


Figure 5: Test scores of PT, PCT, and COPT on WSC, MultiRC, and WiC across T5-Small, T5-Base, and T5-Large.

Figure 5 shows consistent gains with model size. COPT outperforms PCT on WSC and WiC at all scales and achieves higher three-task averages than both baselines. On MultiRC, it trails PCT on Small and Base but slightly surpasses it on Large, indicating that the overall advantage is robust while remaining task-dependent.

5 Conclusion

We revisited common-knowledge extraction and task-specific knowledge learning in multi-source soft prompt transfer, and based on theoretical and practical findings, we propose COPT(a common-knowledge-guided orthogonal prompt transfer method). Experiments show that COPT improves average transfer performance on GLUE and SuperGLUE while retaining a small number of task-specific trainable parameters. The preliminary experiments and ablations further demonstrate the contributions of common-prompt construction and directional constraints on task-specific learning. It is important to note that the theoretical optimality in this paper pertains to the parameter reconstruction risk under given assumptions, and can serve as a feasible approach for theoretical proof, but it is neither the only nor a complete solution. Future work could investigate common-knowledge estimation without exact centering of task-specific offsets, as well as target-dependent source selection and adaptive subspace dimensions, to extend the applicability of multi-source prompt transfer.

AI USE STATEMENT

AI-assisted tools were used for Chinese-to-English translation and language editing; assistance with manuscript drafting, literature and reference organization, and LaTeX preparation; feedback on the theoretical formulation and proofs; assistance with interpreting experimental results; and preparation and revision of plotting scripts and figures. The authors remain responsible for the accuracy, originality, and integrity of the submission, including its claims, proofs, citations, and experimental reporting.

REPRODUCIBILITY STATEMENT

The method section describes common-prompt construction, gradient projection, and target adaptation. The experimental section specifies the datasets, metrics, and training configuration. Appendix A provides proofs and settings for the controlled preliminary experiments, and Appendix B gives computational complexity estimates and measured resource use.

REFERENCES

A. Asai, M. Salehi, M. E. Peters, et al. ATTEMPT: Parameter-Efficient Multi-Task Tuning via Attentional Mixtures of Soft Prompts. In *EMNLP*, pp. 6655–6672, 2022. doi: 10.18653/v1/2022.emnlp-main.446.

R. Belanec, S. Ostermann, I. Srba, et al. Task Prompt Vectors: Effective Initialization through Multi-Task Soft Prompt Transfer. In *ECML PKDD*, pp. 77–94, 2025. doi: 10.1007/978-3-662-72243-5_5.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. In *NeurIPS*, 2023.

C. Eckart and G. Young. The Approximation of One Matrix by Another of Lower Rank. *Psychometrika*, 1:211–218, 1936.

N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-Efficient Transfer Learning for NLP. In *ICML*, pp. 2790–2799, 2019.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2022.

M. Jung, S. Park, J. Sul, et al. Is Prompt Transfer Always Effective? An Empirical Study of Prompt Transfer for Question Answering. In *NAACL-HLT*, pp. 528–539, 2024. doi: 10.18653/v1/2024.naacl-short.44.

B. Lester, R. Al-Rfou, and N. Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In *EMNLP*, pp. 3045–3059, 2021. doi: 10.18653/v1/2021.emnlp-main.243.

J. Y. Li, T. Y. Tang, J. Y. Nie, et al. Learning to Transfer Prompts for Text Generation. In *NAACL-HLT*, pp. 3506–3518, 2022. doi: 10.18653/v1/2022.naacl-main.257.

X. L. Li and P. Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *ACL-IJCNLP*, pp. 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353.

L. Mirsky. Symmetric Gauge Functions and Unitarily Invariant Norms. *Quarterly Journal of Mathematics*, 11:50–59, 1960.

G. H. Qin and J. Eisner. Learning How to Ask: Querying LMs with Mixtures of Soft Prompts. In *NAACL-HLT*, pp. 5203–5212, 2021. doi: 10.18653/v1/2021.naacl-main.410.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Y. S. Su, X. Z. Wang, Y. J. Qin, et al. On Transferability of Prompt Tuning for Natural Language Processing. In *NAACL-HLT*, pp. 3949–3969, 2022. doi: 10.18653/v1/2022.naacl-main.290.

T. Vu, B. Lester, N. Constant, et al. SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer. In *ACL*, pp. 5039–5059, 2022. doi: 10.18653/v1/2022.acl-long.346.

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *NeurIPS*, pp. 3261–3275, 2019a.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *ICLR*, 2019b.

Z. Wang, R. Panda, L. Karlinsky, et al. Multitask Prompt Tuning Enables Parameter-Efficient Transfer Learning. In *ICLR*, 2023. URL https://openreview.net/forum?id=Nk2pDtuhTq.

Y. Wei, A. Tang, L. Shen, Z. Hu, C. Yuan, and X. Cao. Modeling Multi-Task Model Merging as Adaptive Projective Gradient Descent. In *ICML*, 2025.

X. L. Wu, C. L. P. Chen, S. Z. Li, et al. Snapshot Prompt Ensemble for Parameter-Efficient Soft Prompt Transfer. In *ICASSP*, pp. 11946–11950, 2024. doi: 10.1109/ICASSP48485.2024.10448070.

T. Yu et al. Gradient Surgery for Multi-Task Learning. In *NeurIPS*, 2020.

E. B. Zaken, Y. Goldberg, and S. Ravfogel. BitFit: Simple Parameter-Efficient Fine-Tuning for Transformer-Based Masked Language-Models. In *ACL*, pp. 1–9, 2022.

P. Y. Zhang, R. C. Zhang, Z. J. Nie, et al. Dynamic Task Vector Grouping for Efficient Multi-Task Prompt Tuning. In *Findings of ACL*, pp. 26805–26821, 2025. doi: 10.18653/v1/2025.findings-acl.1374.

S. Zhao, S. J. Yang, S. C. Tan, et al. Prompt Contrastive Transformation: An Enhanced Strategy for Efficient Prompt Transfer in Natural Language Processing. *Transactions of the Association for Computational Linguistics*, 13:848–860, 2025. doi: 10.1162/TACL.a.22.

Q. H. Zhong, L. Ding, J. H. Liu, et al. PanDa: Prompt Transfer Meets Knowledge Distillation for Efficient Model Adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 36(9): 4835–4848, 2024. doi: 10.1109/TKDE.2024.3376453.

A Proofs and Supplementary Derivations

A.1 Proof of Theorem 1

By Assumption 3.1, $P_i = C_0 + T_i$ and $\sum_i w_i T_i = 0$. Therefore,

$$\overline{P}_w = \frac{1}{W}\sum_i w_i P_i = C_0 + \frac{1}{W}\sum_i w_i T_i = C_0. \tag{15}$$

For any candidate common matrix C, expanding the squared risk using the Frobenius inner product $\langle A, B\rangle_F = tr(A^\top B)$ gives

$$L(C) = \sum_i w_i \|T_i + (C_0 - C)\|_F^2 \tag{16}$$

$$= \sum_i w_i \|T_i\|_F^2 + W\|C_0 - C\|_F^2 \tag{17}$$

$$+2\langle \sum_i w_i T_i, C_0 - C\rangle_F \tag{18}$$

$$= \sum_i w_i \|T_i\|_F^2 + W\|C_0 - C\|_F^2. \tag{19}$$

The cross term vanishes because $\sum_i w_i T_i = 0$. The first term is independent of C, and $W > 0$, so the two objectives have the same set of minimizers over $\mathcal{C}$. The objective is continuous and has bounded sublevel sets; hence it attains its minimum on the nonempty closed set $\mathcal{C}$. This proves Theorem 1.

A.2 The Unconstrained Common Center

For arbitrary source prompts and fixed positive weights, let $\overline{P}_w = W^{-1}\sum_i w_i P_i$. By the definition of the mean, $\sum_i w_i(P_i - \overline{P}_w) = 0$. Substituting $P_i - C = (P_i - \overline{P}_w) + (\overline{P}_w - C)$ into the squared risk and expanding, the cross term vanishes after weighted summation, yielding

$$L(C) = \sum_i w_i \|P_i - \overline{P}_w\|_F^2 + W\|C - \overline{P}_w\|_F^2. \tag{20}$$

The first term is independent of C, and the second equals zero if and only if $C = \overline{P}_w$. Thus, $\overline{P}_w$ is the unique global minimizer of $L(C)$ over $\mathbb{R}^{m\times d}$. This decomposition holds for any source-prompt collection. Under Assumption 3.1, we further obtain $\overline{P}_w = C_0$.

A.3 Proof of Theorem 2

Appendix A.2 gives the following decomposition for every matrix C:

$$L(C) = \sum_i w_i \|P_i - \overline{P}_w\|_F^2 + W\|C - \overline{P}_w\|_F^2. \tag{21}$$

Write $\overline{P}_w = \sum_{j=1}^q \sigma_j u_j v_j^\top$ with $\sigma_1 \geq \cdots \geq \sigma_q \geq 0$. Therefore, after restricting C to $\mathcal{C}_r$, minimizing $L(C)$ is exactly the classical best rank-r approximation problem for $\overline{P}_w$. By the Eckart–Young–Mirsky theorem (Eckart & Young, 1936; Mirsky, 1960), the minimum approximation error is $\sum_{j=r+1}^q \sigma_j^2$, attained by $C_r^* = \sum_{j=1}^r \sigma_j u_j v_j^\top$. Substituting this value into the decomposition yields $\min_{C\in\mathcal{C}_r} L(C) = \sum_i w_i \|P_i - \overline{P}_w\|_F^2 + W\sum_{j=r+1}^q \sigma_j^2$. Hence C_r^* minimizes the original

Table 3: Task-level results for the common-knowledge extraction and orthogonal-adaptation comparisons (%).

Task	PCT	Theory-PCT	Orthogonal-PCT
MNLI	85.6	85.6	85.7
QQP	90.4	90.3	90.3
QNLI	93.2	93.3	93.3
SST-2	93.1	93.3	93.3
RTE	65.5	80.6	78.4
CoLA	36.9	32.1	49.8
STS-B	89.6	89.1	89.7
MRPC	89.7	88.7	90.7

weighted source-prompt risk. If $\overline{P}_w = C_0$ under Assumption 3.1, Theorem 1 transfers this conclusion to the common-knowledge reconstruction objective. The minimizer is unique when there is a strict singular-value gap at r; otherwise, the truncated SVD specifies one globally optimal solution. This proves Theorem 2.

A.4 Common-Knowledge Extraction: Settings and Task-Level Results

The experiment uses T5-Base as the backbone and GLUE as the evaluation benchmark. The source tasks are MNLI and QNLI, and the target tasks are MNLI, QQP, QNLI, SST-2, RTE, CoLA, STS-B, and MRPC. PCT constructs common knowledge from the first source-task block of the jointly reconstructed matrix, whereas Theory-PCT applies SVD to the mean prompt. The reconstruction rank of the joint SVD is 100. Task-specific knowledge learning uses Adam for 24,000 steps with a learning rate of 0.3 and a batch size of 32. Target-adaptation learning rates are 0.3 for GLUE and 0.1 for SuperGLUE, with 500 warmup steps. Table 3 reports the task-level test metrics for both controlled comparisons.

The orthogonal comparison uses the first eight left singular vectors of the common prompt and transforms source gradients as $\widetilde{g} = (I_m - BB^\top)g$ before optimization. The same training schedule is used for both comparisons, so the merged table supports direct task-level inspection of the two interventions.

B Time Complexity Analysis

We compare the time complexity of three representative methods. Let n be the number of source tasks, $m \times d$ the prompt dimensions, and K the number of source-transformation updates. Source-knowledge preparation starts from pretrained source prompts and excludes their initial training cost. Under $m \leq d$, with dense SVD and the factorized projection $G - B(B^\top G)$, mean computation, decomposition, and source-transformation optimization in COPT yield

$$T^{src}_{COPT} = O(nmd + m^2 d + Knmdr_b). \tag{22}$$

PCT (Zhao et al., 2025) applies SVD to a concatenated matrix of size $nm \times d$ and computes pairwise contrastive losses across source tasks, giving

$$T^{src}_{PCT} = O(nmd \min(nm, d) + Kn^2 md). \tag{23}$$

MPT (Wang et al., 2023) learns a shared prompt through teacher–student distillation. Let S_s denote the number of distillation steps per source task, and let F and B_{bp} denote the costs of backbone forward propagation and backpropagation to prompt parameters at a fixed batch size. Then,

$$T^{src}_{MPT} = O\left(nS_s(F + B_{bp} + md)\right). \tag{24}$$

The constant number of teacher and student forward passes is included in this estimate. With fixed prompt dimensions, rank, and per-task training budgets, the source-dependent leading terms of

Table 4: Measured total transfer time and peak memory of COPT as the source-prompt library grows.

Sources	Target	Seed	Time (s)	Peak memory (MiB)
3	MNLI + QNLI + QQP	RTE	398.663	8300.688
4	MNLI + QNLI + QQP + SST-2	RTE	411.319	8300.969
5	MNLI + QNLI + QQP + SST-2 + SQuAD	RTE	418.837	8301.484
6	MNLI + QNLI + QQP + SST-2 + SQuAD + ReCoRD	RTE	445.430	8302.994

COPT and MPT grow linearly with n. For the current setting, $n = 3, 4, 5, 6$, $m = 100$, and $d = 768$, both PCT's concatenated SVD and its pairwise contrastive term grow quadratically. Figure 4.4(b) represents these normalized growth patterns by $\frac{n}{3}$ and $(n/3)^2$, respectively. Unit computation costs differ across methods, so the figure does not establish an ordering of absolute running times.

During target adaptation, multi-source fusion in COPT and PCT incurs a prompt-level cost of $O(nmd)$, while MPT uses its distilled shared prompt and a rank-one task correction at a cost of $O(md)$. With S_t target updates, including backbone computation gives

$$T^{tgt}_{COPT/PCT} = O\left(S_t(F + B_{bp} + nmd)\right), T^{tgt}_{MPT} = O\left(S_t(F + B_{bp} + md)\right). \tag{25}$$

C Related Work

C.1 Soft Prompt Tuning

Soft prompt tuning is a lightweight approach that prepends a small number of trainable continuous vectors to the input embeddings to guide a frozen pretrained language model toward task-specific outputs (Lester et al., 2021; Li & Liang, 2021). Because adaptation leaves the model architecture and backbone parameters unchanged, each task requires storing only a small set of prompt vectors. This provides low parameter overhead, independent task modules, and convenient reuse (Lester et al., 2021; Zhang et al., 2025).

Moreover, research has shown that an appropriate initialization significantly enhances the parameter performance of prompts (Zhao et al., 2025). Prior work has used discrete prompt search to initialize continuous prompts and provide a more informative starting point, although designing and searching discrete templates generally requires additional human effort or computation (Qin & Eisner, 2021). In low-resource settings, limited supervision may also prevent prompt parameters from converging sufficiently, resulting in substantial variation across random initializations (Wu et al., 2024). Reusing continuous prompts already optimized on other tasks can therefore provide prior knowledge for target-task initialization and ease optimization under limited supervision. Such cross-task reuse or initialization of continuous prompts forms the basis of soft prompt transfer.

C.2 Soft Prompt Transfer

Soft prompt transfer reuses prompts trained on one or more source tasks to provide more informative initialization or structural priors for a target task (Li et al., 2022; Vu et al., 2022; Su et al., 2022; Jung et al., 2024). Early work largely focused on exploring the feasibility of single-source transfer. SPoT first trains soft prompts on a source task or a mixture of tasks and then uses them to initialize the target prompt. It also treats prompt parameters as task representations and retrieves sources likely to yield positive transfer based on prompt similarity (Vu et al., 2022). Su et al. systematically examined prompt transfer across tasks and models, showing that task relatedness, prompt initialization, and model scale all influence transfer gains (Su et al., 2022). However, this approach cannot fully exploit complementary knowledge across multiple source prompts, thereby lacking potential.

To leverage multiple sources, researchers are further exploring approaches based on representation decomposition and prompt ensembling. For example, ATTEMPT constructs a source-prompt library and uses a lightweight attention module to dynamically combine source prompts for each input (Asai et al., 2022). Task Prompt Vectors constructs task vectors from multi-task soft prompts to generate

more effective target-prompt initializations (Belanec et al., 2025). To address redundancy and irrelevant information in multi-source prompts, PCT extracts task-independent embeddings through SVD, separates task-specific embeddings using reconstruction and contrastive objectives, and finally combines them through target-instance attention (Zhao et al., 2025). In summary, existing multi-source prompt transfer methods primarily follow the basic paradigm of knowledge decomposition and target adaptation (Zhang et al., 2025). However, current research still lacks a unified theoretical framework for extracting common knowledge, and the analysis of the relationship between common and specific knowledge remains insufficient.

D Implementation Details

Main experiments use T5-Base (220M parameters) (Raffel et al., 2020); the model-scale analysis additionally uses T5-Small (60M) and T5-Large (770M). Source-prompt training and task-specific knowledge learning follow the hyperparameters reported by PCT (Zhao et al., 2025). During target adaptation, the prompt length is 100, the batch size is 32, weight decay is 1×10^{-5}, and the number of warmup steps is 500. The target-task learning rate is 0.3 for GLUE and 0.1 for SuperGLUE. We use Adam and bfloat16 precision.

COPT first applies truncated SVD to the mean source prompt to construct a common prompt and then uses its left singular vectors to form the shared basis. Unless otherwise specified in the hyperparameter analysis, the common-prompt rank r_p is 25 and the projection-basis dimension r_b is 8.